Journal of Artificial Intelligence Research , Volume 16, 321 -- 357, 2002

SMOTE: Synthetic Minority Over-sampling TEchnique

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer*
Department of Computer Science and Engineering, ENB 118
University of South Florida 4202 E. Fowler Ave.
Tampa, Fl 33620
*Sandia National Laboratories
Biosystems Research Department, P.O. Box 969, MS 9951
Livermore, CA, 94551-0969, USA
Email:{chawla, kwb, hall}@csee.usf.edu, *wpk@ca.sandia.gov

Abstract

This paper describes an approach to dealing with construction of classifiers from imbalanced datasets. A dataset is imbalanced if the classification categories are not approximately equally represented. Usually real-world datasets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples. Often the cost of misclassifying an abnormal (interesting) example as a normal example is much higher than the cost of the reverse error. Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier to the minority class. This paper shows that a combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance than only under-sampling the majority class. Our method of over-sampling the minority class involves creating synthetic minority class examples. Performance is measured using the area under the Receiver Operating Characteristic curve.

1 Introduction

A dataset is imbalanced if the classes are not approximately equally represented. Imbalance on the order of 100 to 1 is prevalent in fraud detection and imbalance of up to 100,000 to 1 has been reported in other applications [Provost and Fawcett, 1997]. There have been attempts to deal with imbalanced datasets in domains such as fraudulent telephone calls [Fawcett and Provost, 1996], telecommunications management [Ezawa et al., 1996], text classification [Lewis and Catlett, 1994] and detection of oil spills in satellite images [Kubat et al., 1998].

Performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and / or the costs difference of errors is large. For example, consider the classification of pixels in mammogram

images for suspiciousness of cancer [Woods et al., 1993]. A typical mammography dataset might contain 98% normal pixels and 2% abnormal pixels. A simple default strategy of guessing the majority class would give a predictive accuracy of 98%. But the nature of the application requires a fairly high rate of correct detection in the minority class and can tolerate a small error rate in the majority class in order to achieve this. Simple predictive accuracy is clearly not appropriate in such situations. The Receiver Operating Characteristic (ROC) curve is a standard technique for summarizing classifier performance over a range of tradeoffs between true positive and false positive error rates [Bradley, 1997] [Swets, 1988]. The Area Under the Curve (AUC) is an accepted performance metric for an ROC curve.

The machine learning community has addressed the issue of class imbalance in two ways. One is to assign distinct costs to training examples [Pazzani et al., 1994] [Domingos, 1999]. The other is to re-sample the original dataset, either by over-sampling the minority class and / or under-sampling the majority class [Kubat and Matwin, 1997] [Japkowicz, 2000] [Lewis and Catlett, 1994] [Ling and Li, 1998]. Our approach blends under-sampling of the majority class with a special form of over-sampling the minority class. Experiments with various datasets and the C4.5 decision tree classifier [Quinlan, 1992] show that our approach improves over other previous re-sampling approaches.

Section 2 gives an overview of performance measures. Section 3 reviews the most closely related work dealing with imbalanced datasets. Section 4 presents the details of our approach. Section 5 presents experimental results comparing our approach to other re-sampling approaches. Section 6 discusses the results and suggests directions of future work.

2 Previous Work: Imbalanced datasets

Provost and Fawcett (1997) have introduced the ROC convex hull method to estimate the classifier performance for imbalanced datasets. They note that the problems of unequal class distribution and unequal error costs are related and that little work has been done to address either problem. In the ROC convex hull method, the ROC space is used to separate classification performance from the class and cost distribution information. The decision goal is projected on to the ROC space which generates a set of iso-performance lines and after that a convex hull is generated. A point on the convex hull intersecting the iso-performance line with the highest True Positive (TP)-intercept will be optimal.

Kubat and Matwin (1997) have looked at under-sampling as a plausible solution. They selectively under-sampled the majority class while keeping the original population of the minority class. They have used the geometric mean as a performance measure for the classifier, which can be related to a single point on the ROC curve. The minority examples were divided into four categories: some noise overlapping the positive class decision region, borderline samples, redundant samples and safe samples. The borderline examples were detected using the Tomek links concept [Tomek, 1976]. Another related work, [Kubat et al., 1998] proposed the SHRINK system that classifies an overlapping region of minority and majority classes as positive; it searches for the “best positive region”.

Japkowicz (2000) has discussed the effect of imbalance in the dataset. She has evaluated

three strategies: under-sampling, resampling and a recognition based induction scheme. We focus on the sampling approaches. She experimented on artificial 1D data in order to easily measure and construct concept complexity. Two resampling methods were considered. Random resampling consisted of resampling the smaller class at random until it consisted of as many samples as the majority class and "focused resampling" consisted of resampling only those minority examples that occurred on the boundary of minority and majority. Random under-sampling was considered, which involved under-sampling the majority class at random until it matched the minority class. Focused under-sampling involved under-sampling the majority class samples lying further away. It was noted that both the sampling approaches were effective and using the sophisticated sampling techniques didn't give any clear advantage in the domain considered.

One approach that is relevant to our work is that of Ling and Li (1998). They combine over-sampling of the minority class with under-sampling of the majority class. They used lift analysis instead of accuracy to measure a classifier's performance. They proposed that the testing examples be ranked by a confidence measurement and then lift be used as the evaluation criteria. A lift curve is similar to ROC curve, but is more tailored for the marketing analysis problem [Ling and Li, 1998]. Ada-boosted C4.5 and ada-boosted Naive Bayes were the learning algorithms used for experiments. In one experiment, they under-sampled the majority class and noted that the best lift index is obtained when the classes are equally represented [Ling and Li, 1998]. In another experiment, they over-sampled the positive (minority) examples with replacement to match the number of negative (majority) examples to the number of positive examples. The over-sampling and under-sampling combination didn't get significant improvement in the lift index.

Lewis and Catlett (1994) examined heterogeneous uncertainty sampling for supervised learning. This method is useful for training samples with uncertain classes. The training samples are labeled incrementally in two phases and the uncertain instances are passed on to the next phase. They modified C4.5 to include a loss ratio for determining the class values at the leaves. The class values were determined by comparison with a probability threshold of $LR/(LR+1)$.

Another approach that is similar to our work is that of Domingos (1999). He compares the "metacost" approach to each of majority under-sampling and minority over-sampling. He finds that metacost improves over either, and that under-sampling is preferable to minority over-sampling. Error-based classifiers are made cost-sensitive. The probability of each class for each example is estimated and the examples are then relabeled with the optimal class by a cost matrix. The relabeling of the examples expands the decision space as it creates new samples for the classifier to learn on.

To summarize the experience of the literature, under-sampling the majority class performs better than over-sampling the minority class. A combination of the two as done in previous work does not outperform only under-sampling. But, the over-sampling of the minority class has been done by sampling with replacement from the original data. Our approach uses a different method of over-sampling.

3 SMOTE: Synthetic Minority Over-sampling TEchnique

3.1 Minority over-sampling with replacement

The approach of over-sampling the minority class by sampling with replacement is seemingly attractive. However, [Ling and Li, 1998] [Japkowicz, 2000] have discussed over-sampling with replacement and have noted that there isn't a significant improvement in minority class recognition with it. We interpret the underlying effect in terms of decision regions in feature space. Essentially, as the minority class is over-sampled by increasing amounts, the effect, using decision trees, is to identify similar but more specific regions in the feature space as the decision region for the minority class. This effect can be understood from the plots in Figures 1, 2 and 3.

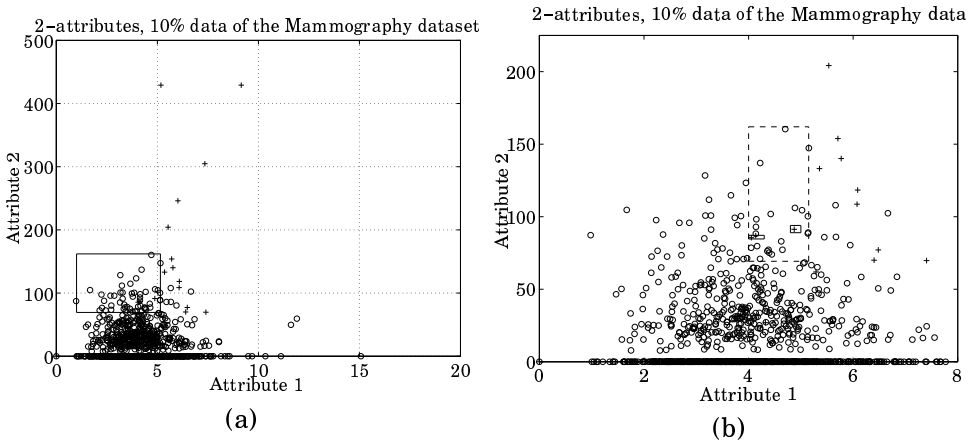

Figure 1: a) Decision region the three minority class samples (shown by '+') come under. The decision region is indicated by the solid-line rectangle. b) A zoomed-in view of the chosen minority class samples for the same dataset. Dashed lines show the decision region after over-sampling the minority with synthetic generation and the small solid-line rectangles show the effect of over-sampling the minority with replication.

The data for the plot in Figure 1 was extracted from a Mammography dataset [Woods et al., 1993][1]. The minority class samples are shown by + and the majority class samples are shown by *o* in the plot. In Figure 1(a), the region indicated by the solid-line rectangle is a majority class decision region. Nevertheless, it contains three minority class samples shown by '+' as false negatives. If we replicate the minority class, the decision region for the minority class becomes very specific and will cause splits in the decision tree. This will lead to more terminal nodes (leaves) as the learning algorithm tries to learn more and more specific regions of the minority class; in essence, overfitting. Replication of the minority class does not cause its decision boundary to spread into the majority class region. Thus, in Figure 1(b), the three samples previously in the majority class decision region now carve very specific decision regions around them.

[1]The data is available from USF Intelligent Systems Lab, http://morden.csee.usf.edu/∼chawla

3.2 SMOTE

We propose an over-sampling approach that over-samples the minority class by creating "synthetic" examples rather than by over-sampling with replacement. This approach is motivated by a technique that proved successful in handwritten character recognition [Thien and Bunke, 1997]. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, the neighbors from the k nearest neighbors are randomly chosen. Our implementation currently uses five nearest neighbors. For instance, if the amount of over-sampling needed is 200%, only two neighbors from the five nearest neighbors are chosen and one sample is generated in the direction of each. Synthetic samples are generated by taking the difference between the feature vector (sample) under consideration and its nearest neighbor. This difference is multiplied by a random number between 0 and 1, and added to the feature vector under consideration. This causes selection of a random point along the line segment between two feature points. This approach effectively forces the decision region of the minority class to become more general.

The synthetic examples push the classifier to create larger and less specific decision regions as shown by the dashed lines in Figure 1(b), rather than smaller and more specific regions. More general regions are now learned for the minority class samples rather than them being subsumed by the majority class samples around them. The effect is that decision trees generalize better on unseen cases. Figures 2 and 3 compare the minority over-sampling with replacement and SMOTE. The experiments were conducted on the mammography dataset. There were 10923 examples in the majority class and 260 examples in the minority class originally. We had approximately 9831 examples in the majority class and 233 examples in the minority class for the training set used in 10-fold cross-validation. There is a reduction in the sizes of the training and test sets as the original data was separated into 90% for traiing and 10% for testing. The minority class was over-sampled at 100%, 200%, 300%, 400% and 500% of its original size. The performance and the decision tree sizes are averages over 10-fold cross-validation. The graphs show that the tree size for minority over-sampling with replacement is much greater than that for SMOTE, and the minority class recognition performance of the minority over-sampling with replacement for higher degrees of replication doesn't increase as much as SMOTE.

3.3 Under-sampling and SMOTE Combination

The majority class is under-sampled by randomly removing samples from the majority population. Under-sampling of the majority class is done so that the minority class becomes some percent of the majority class. This forces the learner to experience varying degrees of under-sampling and at higher degrees of under-sampling the minority class has a larger presence in the training set. For instance, under-sampling the majority class at 200% means that the modified dataset will contain twice as many samples from the minority class as from the majority class; that is, if the minority class has 50 samples and the majority class has 200 samples and we under-sample the majority class at 200%, the majority class would end up having 25 samples. By applying a combination of under-sampling and over-sampling, the initial bias of the learner towards the negative (majority)

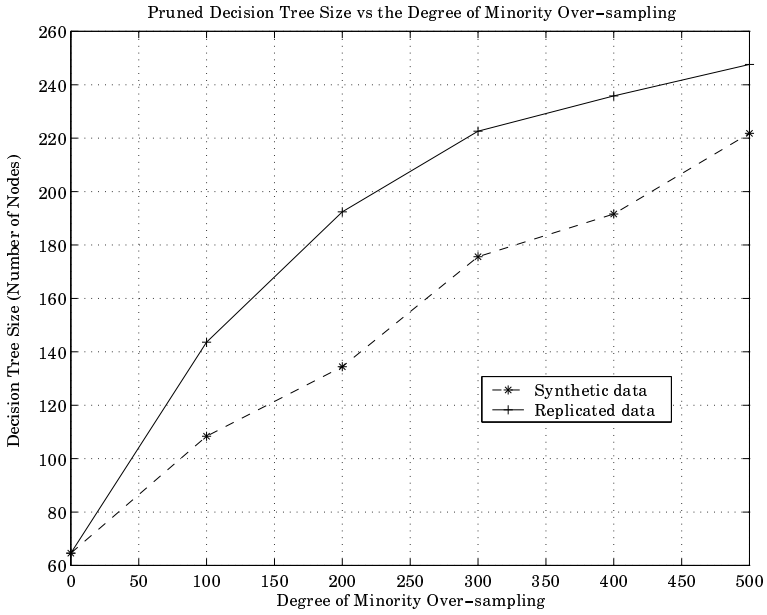

Figure 2: Comparison of decision tree sizes for replicated over-sampling and SMOTE

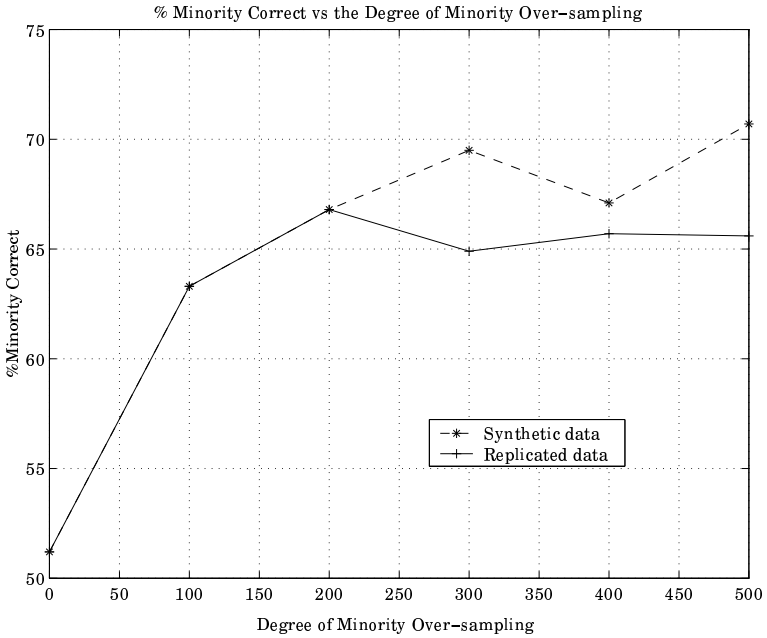

Figure 3: Comparison of % Minority correct for replicated over-sampling and SMOTE

class is reversed in the favor of the positive (minority) class. Decision trees are learned on the dataset perturbed by "SMOTING" the minority class and under-sampling the majority class.

4 Experiments

We compared various combinations of SMOTE and under-sampling against plain under-sampling. We used C4.5 release 8 [Quinlan, 1992] as the base classifier. ROC curves were plotted by taking %FP on the X-axis and %TP on the Y-axis. %FP and %TP were averaged over the 10-fold cross-validation runs for each of the data combinations. The minority class examples were over-sampled by calculating the five nearest neighbors and generating synthetic examples. The AUC was calculated using the trapezoidal rule. We extrapolated an extra point of TP = 100% and FP = 100% for each ROC curve.

4.1 Datasets

We experimented on six different datasets. In order of increasing imbalance they are:

1. Pima Indian Diabetes [Blake and Merz, 1998] has 2 classes and 768 samples. The data is used to identify the positive diabetes cases in a population near Phoenix, Arizona. The number of positive class samples is only 268. The sensitivity to detection of diabetes cases will be a desirable attribute of the classifier.

2. Phoneme dataset is from the ELENA project [2]. The aim of the dataset is to distinguish between nasal (class 0) and oral sounds (class 1). There are 5 attributes. The class distribution is 3818 samples in class 0 and 1586 samples in class 1.

3. Satimage dataset [Blake and Merz, 1998] has 6 classes originally. We chose the smallest class as the minority class and collapsed the rest of the classes into one [Provost et al., 1998]. This gave us a skewed 2-class dataset, with 5809 majority class samples and 626 minority class samples.

4. The Forest cover dataset from UCI repository [Blake and Merz, 1998]. This dataset has 7 classes and 581,012 samples. This dataset is for the prediction of forest cover type based on cartographic variables. Since our system only works for binary classes[3] we extracted two classes data from this dataset and ignored the rest. The two classes we considered are Ponderosa Pine with 35754 samples and Cottonwood/Willow with 2747 samples.

5. The Oil dataset was provided by Robert Holte and is used in their paper [Kubat et al., 1998]. This dataset has 41 oil slick samples and 896 non-oil slick samples.

6. The Mammography dataset [Woods et al., 1993] has 11183 samples with 260 calcifications. If we look at predictive accuracy as a measure of goodness of the classifier

[2] ftp.dice.ucl.ac.be in the directory pub/neural-nets/ELENA/databases.

[3] Most other approaches only work for two classes,[Ling and Li, 1998] [Japkowicz, 2000] [Kubat and Matwin, 1997] [Provost and Fawcett, 1997]

for this case, the default accuracy would be 97.68% that is every sample is labeled non-calcification. But, it is desirable for the classifier to predict more of calcifications correctly.

4.2 ROC Creation

An ROC curve is produced by using C4.5 to create a classifier for each one of a series of modified training datasets. A given ROC curve is produced by first over-sampling the minority class to a specified degree and then under-sampling the majority class at increasing degrees to generate the successive points on the curve. Different ROC curves are produced by starting with different levels of minority over-sampling.

Figures 4 and 5 show the experimental ROC graphs for two of the datasets discussed earlier. In each Figure, the ROC curve for simple under-sampling of the majority class [Ling and Li, 1998] [Japkowicz, 2000] [Kubat and Matwin, 1997] [Provost and Fawcett, 1997] is compared with our approach of combining synthetic minority over-sampling (SMOTE) and majority under-sampling. The simple under-sampling curve is labeled with "under". Depending on the size and relative imbalance of the dataset, one to five SMOTE and under-sampling (over and under) curves are plotted.

We only show the winner SMOTE and under-sampling combination and the simple under-sampling curve in the graphs. Each point on the ROC curve is a classifier learned for a particular combination of under-sampling and over-sampling. The lower leftmost point for a given ROC curve is the dataset without any majority class under-sampling and minority class over-sampling, that is the performance of C4.5 on the original dataset.

For instance, in the set of ROC curves for the phoneme dataset, Figure 4, there are two ROC curves. One is for simple under-sampling in which the range of under-sampling is varied between 5% and 2000% at different intervals. The other curve is for the minority class over-sampling with synthetic generation technique (SMOTE). The ROC shown is for the minority class over-sampled at 200%. Each point on the SMOTE ROC curves represents a combination of over-sampling and under-sampling, the amount of under-sampling follows the same range as laid down for the simple under-sampling technique. Table 1 lists the AUCs for the ROC curves of all the datasets. The convention used in the table and graphs is:

Under $->$ The majority class is under-sampled so that the majority class is some proportion of minority class but the minority class is not over-sampled.

x OU $->$ The minority class is over-sampled by x% and the majority class is under-sampled so that the majority class is some proportion of minority class.

The ROC curves show a trend that as we increase the amount of under-sampling coupled with over-sampling, our minority classification accuracy increases, of course at the expense of more majority class errors.

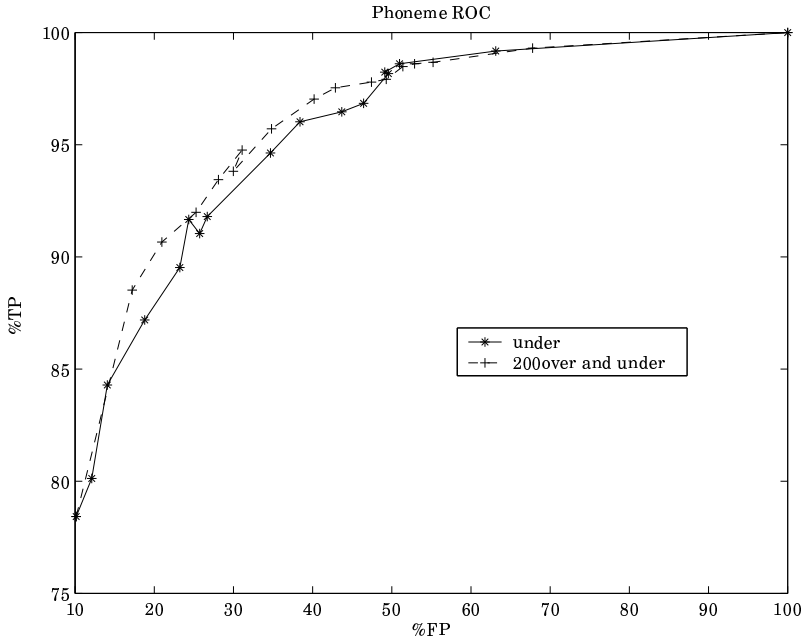

Figure 4: Phoneme

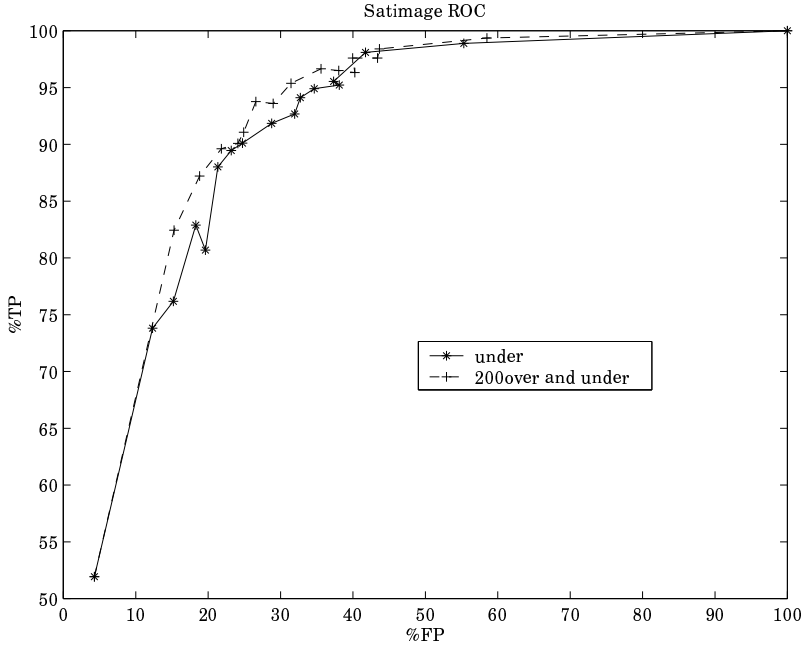

Figure 5: Satimage

Table 1: AUC's with the best highlighted in bold

Dataset	Under	100 OU	200 OU	300 OU	400 OU	500 OU
Pima	7241.98	**7301.17**				
Phoneme	8621.77	8643.99	**8661.38**			
Satimage	8900.21	8956.69	**8979.08**	8962.86	8974.83	8960.05
Forest Cover	9807.35	9832.37	9834.18	**9849.59**	9841.19	9842.44
Mammography	9300	9264.04	9261.5	9316.53	**9332.03**	9302.23
Oil	8524.4	8523.24	8368.03	8161.48	8339.08	**8537.11**

4.3 AUC Calculation

The Area Under the ROC curve (AUC) is calculated using a form of the trapezoid rule. The lower leftmost point for a given ROC curve is the C4.5 performance on the raw data. The upper rightmost point is always (100%, 100%). If the curve does not naturally end at this point, the point is added. This is necessary in order for the AUC's to be compared over the same range of %FP.

The AUCs listed in the Table 1 show that for all datasets, the combined synthetic minority oversampling and majority oversampling is able to improve over plain majority oversampling. Thus, our SMOTE approach provides an improvement in sensitivity to correct classification of data in the underrepresented class.

5 Summary and Discussion

The results show that the SMOTE approach holds a lot of promise. The SMOTE approach provides a new definition for over-sampling. The combination of SMOTE and under-sampling performs better than plain under-sampling. The SMOTE approach was tested on a variety of datasets, ranging in the degrees of imbalance and even the amount of data in the training set, thus giving a diverse test bed. One limitation of SMOTE is that it is only applicable for binary class problems with a continuous feature space. It forces focused learning and introduces a learning bias towards the minority class. Usually, the minority class is the positive class and the target that the classifier would want to maximize its performance on.

The interpretation of why synthetic minority oversampling improves performance whereas minority oversampling with replacement does not is actually fairly straightforward. Consider the effect on the decision regions in feature space when minority oversampling is done by replication (sampling with replacement) versus introduction of synthetic examples. With replication, the decision region that results in a classification decision for the minority class can actually become smaller and more specific as the minority samples in the region are replicated. This is the opposite of the desired effect. Our method of synthetic over-sampling works to cause the classifier to build larger decision regions that contain nearby minority class points.

There are many topics to be considered further in this line of research. Automated adaptive selection of the number of nearest neighbors would be valuable. Different strategies for

creating the synthetic neighbors may be able to improve the performance. Also, selecting nearest neighbors with a focus on examples that are incorrectly classified may improve performance.

Acknowledgments

This research was partially supported by the United States Department of Energy through the Sandia National Laboratories ASCI VIEWS Data Discovery Program, contract number DE-AC04-76DO00789.

References

[Blake and Merz, 1998]
C.L. Blake and C.J. Merz. UCI Repository of Machine Learning Databases http://www.ics.uci.edu/~mlearn/ MLRepository.html, University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[Bradley, 1997]
A. P. Bradley. The use of the Area Under the ROC Curve in the evaluation of Machine Learning Algorithms. *Pattern Recognition*, Vol. 30(6), pp. 1145–1159, 1997.

[Domingos, 1999]
P. Domingos. Metacost: A General Method for Making Classifiers Cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, 1999.

[Ezawa et al., 1996]
J. Ezawa, K., M. Singh, and W. Norton, S. Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management. In *Proceedings of the International Conference on Machine Learning, ICML-96*, pp. 139–147, Bari, Italy, 1996. Morgan Kauffman.

[Fawcett and Provost, 1996]
T. Fawcett and F. Provost. Combining Data Mining and Machine Learning for Effective User Profile. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 8–13, Portland, OR, 1996. AAAI.

[Japkowicz, 2000]
N. Japkowicz. The Class Imbalance Problem: Significance and Strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning*, 2000.

[Kubat and Matwin, 1997]
M. Kubat and S. Matwin. Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179–186, 1997.

[Kubat et al., 1998]
M. Kubat, R. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, Vol. 30, pp. 195–215, 1998.

[Lewis and Catlett, 1994]
D. Lewis and J. Catlett. Heterogeneous Uncertainity Sampling for Supervised Learning. In *Proceedings of the Eleventh International Conference of Machine Learning*, pp. 148–156, 1994.

[Ling and Li, 1998]
C. Ling and C. Li. Data Mining for Direct Marketing Problems and Solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, 1998.

[Pazzani et al., 1994]
M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. Reducing Misclassification Costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, 1994.

[Provost and Fawcett, 1997]
F. Provost and T. Fawcett. Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pp. 43–48, 1997.

[Provost et al., 1998]
F. Provost, T. Fawcett, and R. Kohavi. The Case Against Accuracy Estimation for Comparing Induction Algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 445–453, 1998.

[Quinlan, 1992]
J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1992.

[Swets, 1988]
J. Swets. Measuring the Accuracy of Diagnostic Systems. In *Science*, pp. 1285–1293. 1988.

[Thien and Bunke, 1997]
M. Thien and H Bunke. Off-line, Handwritten Numeral Recognition by Perturbation method. *Pattern Analysis and Machine Intelligence*, Vol. 19/5, pp. 535–539, 1997.

[Tomek, 1976]
I. Tomek. Two modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 6, pp. 769–772, 1976.

[Woods et al., 1993]
K. Woods, C. Doss, K. Bowyer, J. Solka, C. Priebe, and W. Kegelmeyer. Comparative evaluation of Pattern Recognition techniques for detection of microcalcifications in Mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 7(6), pp. 1417–1436, 1993.